Neurocomputing 640 (2025) 130316

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Relevance-aware prompt-tuning method for multimodal social entity and relation extraction

Zhenbin Chen [a,b], Zhixin Li [a,b,*], Mingqi Liu [a,b], Canlong Zhang [a,b], Huifang Ma [c]

[a] Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
[b] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
[c] College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China

ARTICLE INFO

Communicated by M. Gallo

Keywords:
Named entity recognition
Relation extraction
Multi-modal learning
Prompt-tuning
Dynamic router mechanism

ABSTRACT

Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE) aim to identify specific entities from given text–image pairs and classify the semantic relationships between them, and they have significant applications in social media platform analysis. However, the images and text in social media data are not always aligned, which makes the existing multimodal entity and relation extraction methods still mainly rely on text information. And those mismatched images can even introduce modality noise, leading to negative impacts on the model and preventing them from achieving better performance. To solve this issue, we propose a Relevance-Aware Prompt-tuning (RAP) method with dynamic router mechanism for multi-modal entity and relation extraction. Our method can adaptively learn effective multimodal features from various types of information as prompt vectors and utilize prompt-tuning for entity and relation extraction. Additionally, when integrating information from different modalities, we take into account the intermodal relevance to reduce the negative impact of mismatched visual information on the model, which allows our model to overcome modality noise and achieve better performance. Extensive experiments on three benchmark datasets of tweets demonstrated the effectiveness and superiority of our proposed approach, and achieved approximately 2% increase in F1 values on the three datasets, respectively.

1. Introduction

Social media platforms such as Twitter, WeChat, and Weibo have become crucial channels for information dissemination, where users can express their views and attitudes towards societal events. Extracting information from social media data is of significant importance for understanding public opinion trends and government decision-making. Entity and relation extraction, as fundamental tasks in information extraction, have been widely applied in social media text analysis and have yielded satisfactory results. Named entity recognition (NER) [1–3] task aims to identify named entities like person, location, organization, time, biological protein, etc. in text. While relation extraction (RE) [4–9] attempts to identify potential semantic relationships between entities and outputs the relation triplets. A relation triplet consists of two entities and a relation between them, which is very useful for many natural language processing tasks such as machine reading comprehension [10], machine translation [11], abstractive summarization [12,13], etc. Although many entity and relation extraction methods have been applied to information extraction from social media, their effectiveness is not satisfactory. The main reason for this phenomenon is that social media texts typically have length limitations, usually around 140 characters, which results in many social media texts lacking complete context.

With the advancement of social media technology and multimodal learning [14–16], people are gradually starting to express their opinions on social platforms using images and videos. The development of multimodal learning has also demonstrated the ability of different modalities of information to synergistically collaborate. Therefore, the multi-modal named entity recognition (MNER) [17–21] and multi-modal relation extraction (MRE) [22–27] tasks for multi-modal social media data have been proposed and attracted the attention of researchers. Social images can provide additional contextual information to alleviate the semantic sparsity issue of social text. However, multi-modal entity and relation extraction still suffer from the influence of **modality noise**. As shown in Fig. 1, **mismatched image–text pairs** have generated similar negative impacts on both MNER and MRE tasks. In Fig. 1(a), the mismatch arises when the image depicts Hilton hotels, leading the model to incorrectly identify the entity type of "Konrad

* Corresponding author at: Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China.
E-mail addresses: chenzb@stu.gxnu.edu.cn (Z. Chen), lizx@gxnu.edu.cn (Z. Li), Liumq@stu.gxnu.edu.cn (M. Liu), clzhang@gxnu.edu.cn (C. Zhang), mahuifang@nwnu.edu.cn (H. Ma).

https://doi.org/10.1016/j.neucom.2025.130316
Received 8 May 2024; Received in revised form 28 February 2025; Accepted 18 April 2025
Available online 7 May 2025

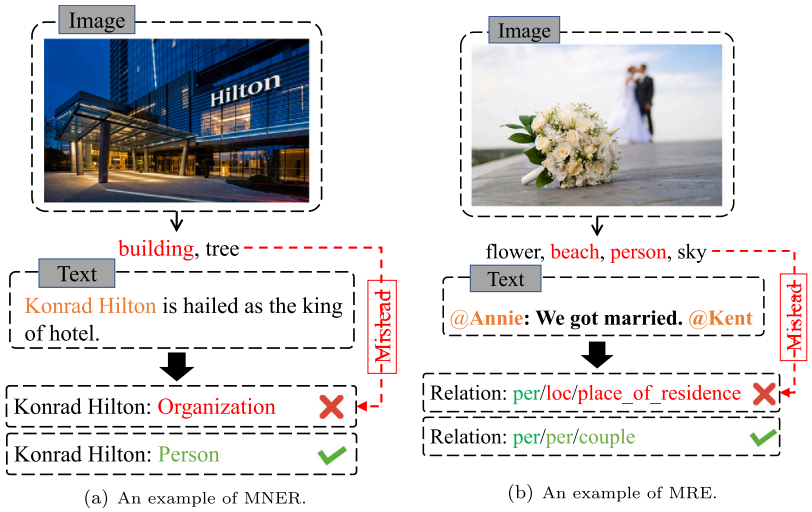


Fig. 1. Modality noise present in MNER and MRE tasks.

Hilton" in the text as a company or institution. However, "Konrad Hilton" refers to the creator of Hilton hotels, and its correct type should be PERSON. Similarly, in Fig. 1(b), the highest confidence in the image is attributed to scenes such as flowers and beaches, while the characters in the image are blurred. This misleads the model into making erroneous judgments about the entities. Therefore, mitigating the issue of modal noise effectively is crucial for both MNER and MRE tasks.

On the other hand, **treating all samples impartially** is also one of the reasons contributing to modal noise. When the social text already provides the necessary contextual semantics to determine entity types and their relationships, visual information becomes unnecessary. We define such samples as simple samples, whereas samples that lack explicit trigger words in the context and require the integration of information from different modalities for inference, we define as hard samples. For the easy samples, the model can accurately extract the entity type and relationship based on textual features alone, without the need for visual information such as the example illustrated in Fig. 1(a). The model can infer the relation between entities as "couple" based on the word "married". However, visual information instead misled the model, preventing correct classification of this easy sample. If the model can dynamically integrate visual information based on the needs of the text, it can further alleviate the problem of modal noise.

To solve these issues, we propose a Relevance-Aware Prompt-tuning (RAP) method with dynamic router mechanism for multi-modal entity and relation extraction. To address the issue of mismatch between image and text, we propose a relevant graph reasoning module to calculate the correlation coefficients of image-text pairs. When integrating features from different modalities, these correlation coefficients help mitigate the negative impact of mismatched visual features on the model. On the other hand, we use a dynamic routing mechanism that enables the model to adaptively learn different feature fusion paths for various samples, as shown in Fig. 2. Furthermore, we discover that prompt fine-tuning methods can better prevent the negative effects of modal noise on the model. Therefore, we utilize multimodal features to compute prompt vectors and employ prompt-tuning methods to extract entities and relationships.

Our main contributions are as follows:

- We propose a dynamic routing mechanism that allows each sample to adaptively select an appropriate feature extraction strategy, which makes the model more efficient and mitigates the impact of modality noise.
- To evaluate the semantic consistency and alignment between the text and images in social media posts, we modeled the relevance representations between modalities and employ graph reasoning algorithm to calculate their alignment. Based on the calculated modality correlation, our model will weight the visual information to prevent irrelevant visual information from being considered.
- We propose to apply prompt-tuning methods to MNER and MRE tasks, and calculate prompt vectors based on joint modality features to further mitigate the impact of modality noise on model decisions.

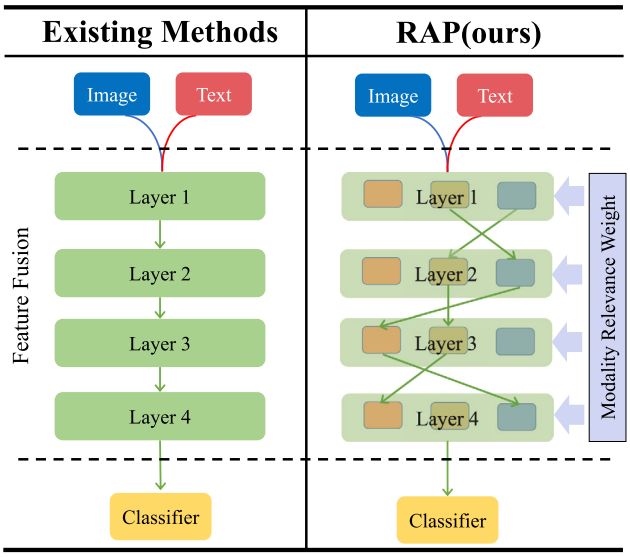


Fig. 2. The core idea of RAP. In the model architecture of RAP(ours), squares of different colors in each layer represent units of different fusion functions.

Extensive experiments conducted on three benchmark datasets, namely Twitter2015, Twitter2017, and MNRE, demonstrate the superiority of our approach, achieving state-of-the-art performance.

2. Related work

2.1. Traditional entity and relation extraction

Traditional entity and relation extraction tasks are important subtasks in the field of information extraction, but they solely focus on extracting from the text modality. They leverage semantic features of the text to mine valuable information from the context to

infer entity types and relationships. Early research on Named Entity Recognition and Relation Extraction primarily focused on hand-crafted feature engineering and extraction templates, which required a significant amount of human labor [28–32]. With the advancement of word embeddings and deep learning techniques, we no longer need to construct the hand-crafted text features. Through pre-trained language models like Word2Vec [33] and Glove [34], we can easily obtain word embeddings with stronger semantic expressiveness. Deep learning models can leverage these word embeddings to better extract semantic information from context for information extraction tasks. However, for entity and relation extraction tasks, the positional information of entities is crucial. Therefore, Zhang et al. [35] first proposed using bidirectional LSTM to model the complete sequential information of all words in a sentence. Additionally, they utilized dependency parsing and WordNet to construct semantic features for relation extraction tasks. Similarly, Lample et al. [36] construct and label segments using a transition-based approach inspired by shift-reduce parsers and utilize bidirectional LSTMs and conditional random fields for NER, which is the first NER method that does not rely on hand-crafted features and domain-specific knowledge. Subsequently, based on this paradigm, many excellent NER and RE tasks have been proposed. [37–42].

With the development of language models, pretrained language models have begun to be pretrained on large-scale corpora, endowing them with stronger natural language understanding capabilities. By fine-tuning pretrained language models using appropriately designed neural networks, we can rely on the powerful NLU capabilities of PLMs for entity and relation extraction. Therefore, Yu et al. [43] used BERT [44] to extract text embeddings and introduced graph-based dependency parsing to provide global features for nested named entity recognition. Additionally, Li et al. [45] innovatively reformulated the NER task as a Machine Reading Comprehension (MRC) task, and they used BERT to extract text embeddings to construct a unified extraction framework. As for relation extraction, SpanBERT [6], KnowBERT [46], and LUKE [7] have become important baselines. SpanBERT [6] restructured the relation extraction task as a span prediction task, improving the span prediction capabilities of pretrained language models by masking contiguous random spans and training span boundary representations. KnowBERT [46] proposed a general method to embed multiple knowledge bases (KBs) into large-scale models, obtaining structured enhanced representations for relationship extraction tasks. Of course, there are many other works using pretrained language models that have achieved satisfactory results [5,7,47,48].

With the popularity of social media platforms [19,25,49–51], extracting valuable information from massive data for news aggregation, natural disaster monitoring, and sentiment analysis [52–55] has become challenging. Therefore, entity and relationship extraction algorithms are widely used for information extraction from social media and have shown good results. However, due to the limitations of social platforms, the length of social text is restricted, leading to a lack of sufficient context information for text-based NER and RE methods. Multimodal entity and relationship extraction have been proposed to address these issues.

2.2. Multimodal entity and relation extraction

2.2.1. Multimodal named entity recognition

The objective of the Multimodal Named Entity Recognition (MNER) task is to identify and categorize entity mentions within text, while also leveraging visual cues from accompanying images. Early studies [17,21,56] have emphasized the importance of incorporating visual features to enhance the learning of representations. Arshad et al. [18] introduced a Gated Fusion mechanism based on semantic alignment scores to weight visual clues and reduce the impact of mismatched visual information. Zhao et al. [57] used a span detection subtask to obtain entity representations as a bridge between the two modalities. Subsequent research [18,58] has identified that incorrect entity recognitions often stem from irrelevant images that disrupt visual attention and misdirect the identification process. To address this problem, Yu et al. [58] introduced the Unified Multimodal Transformer (UMT), which employs a visual gate to adaptively integrate visual data into the final representation. Furthermore, approaches like RIVA [59] and RpBERT [60] have been developed to explicitly assess the relevance between the image and text, treating the binary classification of their relationship as a supportive task, thereby mitigating the confusion caused by unrelated visual inputs.

In more recent advancements, ITA [61] leverages the objects present in the image and combines them with the full text to achieve a cohesive representation and to refine the attention mechanism over textual data. MNER-QG [62] reconceptualizes the MNER task as a machine reading comprehension challenge, prompting models to query for entity recognitions. MoRe [63] augments the MNER model by incorporating contextually relevant texts retrieved through information retrieval methods. UMGF [64] addresses these problems by capturing the visual information of entities from the corresponding image region, preventing irrelevant visual targets from impacting the model. And GEI [57] utilizes a heterogeneous graph interaction network to allow entities to interact with object nodes, capturing visually relevant information and eliminating visual noise associated with non-entity labels.

2.2.2. Multimodal relation extraction

Multimodal Relation Extraction (MRE) aims to use visual information as an auxiliary to better understand the relationships between entities mentioned in the text. Research on MRE is relatively scarce, primarily due to the lack of high-quality multimodal relation extraction datasets. It was not until Zheng et al. [25] introduced the first multimodal relation extraction dataset that MRE research began to gain attention. However, due to severe modality noise issues in the MRE datasets, current MRE methods have certain limitations, which may be one of the reasons for the relatively limited research on MRE. Later, Zheng et al. [26] proposed MEGA, which utilizes a graph structure alignment method to capture the relationships between visual objects and text entities, enabling more accurate identification of relationships between entities in the text. Chen et al. [24] proposed HVPNeT, which enhances the representation of visual information by applying a feature pyramid network to the visual data and employs dynamic gating to facilitate interaction between visual and text features at different layers, leading to better semantic alignment. Although the aforementioned methods have achieved good results, HVPNeT is the first MRE method to consider visual noise. It improves MRE performance by using dynamic gates to filter out visual objects irrelevant to the text. Therefore, subsequent works have gradually focused on addressing the modality noise problem in MRE. For instance, Cui et al. [65] proposed the MIMB model, which uses a Refinement Regularizer designed based on the information bottleneck principle to retain task-relevant information while reducing noise unrelated to the task from each modality, thereby improving model performance. Wang et al. [66] proposed PromptMNER, which utilizes a prompt learning approach to extract image cues related to entities and better integrates information from different modalities through a modality-aware attention mechanism. In addition to the above works, some studies have attempted to improve multimodal relation extraction performance in other ways. For example, Hu et al. [67] enhanced the performance of multimodal entity and relation extraction tasks through innovative pre-training objectives and a soft pseudo-label generation strategy. Hu et al. [68] improved cross-modal reasoning by retrieving visual and textual evidence related to objects, sentences, and the entire image. Multimodal relation extraction is still in a phase of rapid development, and many research problems remain to be explored in the future.

2.3. Prompt tuning

With the development of pre-trained language model technology, the fine-tuning paradigm has become widely adopted in natural language processing tasks. This involves adding a carefully designed classifier on top of the pre-trained language model, enabling the model's powerful natural language understanding capabilities to be applied to downstream tasks. It has been proven that this paradigm is effective and has been widely used across various natural language processing tasks. However, this fine-tuning approach forces the model trained under pre-training objectives to adapt to downstream tasks, and when the gap between the pre-training and downstream tasks is large, the capabilities of the pre-trained language model may not be fully utilized. Therefore, in recent years, researchers have started to explore a more flexible fine-tuning paradigm—Prompt Tuning [69,70]. The core idea of Prompt Tuning is to design different prompt templates for different downstream tasks, transforming these tasks into pre-training tasks (such as the Masked Language Model, MLM), thus allowing the pre-trained language model to overcome this bottleneck. One representative work is PPT [71], which unifies different downstream tasks into a few formats and designs self-supervised pre-training tasks for each format. This enables the model to achieve or surpass full-model fine-tuning performance by adjusting only the soft prompts with a small number of samples. Later, researchers realized that the prompt templates do not necessarily have to be in natural language, nor do they need to be human-understandable; they can be continuous representations. As a result, Li et al. [72] proposed Prefix Tuning, which inserts a learnable vector as a prompt into the pre-trained language model, allowing it to achieve better results than discrete prompts across various natural language processing tasks. Later, Liu et al. [73] discovered that for pre-trained language models with a large number of layers, Prefix Tuning did not perform well. The main reason is that during multi-layer forward propagation, the information in the input layer may be diluted or lost, causing the prompt to lose its effect. To address this, they proposed P-tuning v2, which concatenates corresponding prompt vectors at each layer of the pre-trained language model, effectively solving the problem. These methods not only achieve good results in NLP tasks but also perform well in multimodal tasks [74] and visual-language model pre-training tasks [75,76]. In our approach, we also utilize P-tuning v2 and further improve upon it, achieving better results in multimodal entity and relation extraction tasks.

3. Methods

In this section, we will begin to introduce the Relevance-Aware Prompt-tuning (RAP) method with dynamic router mechanism for multi-modal entity and relation extraction. Firstly, we will introduce how to obtain different fine-grained modal features from text and images for MNER and MRE tasks in Section 3.2. And then, we will introduce the calculation of relevance representation and the graph inference algorithm used for computing modality correlations in Section 3.3. Next, we will employ the dynamic routing mechanism to obtain joint modality features in Section 3.4. Lastly, in Section 3.6, we will use the acquired joint modality features in combination with prompt learning for entity and relationship extraction. The overall framework of our proposed modal is depicted in Fig. 3.

3.1. Task definition

3.1.1. Multi-modal named entity recognition

For a given sentence T and its corresponding image I, which are derived from the same social media post, the goal of NER is to identify the named entities contained within the sentence T and assign them to predefined entity types. Specifically, let $T = (t_1, t_2, \ldots, t_{n-1}, t_n)$ represent an input sentence consisting of n tokens. Our objective is to predict the corresponding sequence labeling $Y = (y_1, y_2, \ldots, y_{n-1}, y_n)$ by incorporating the image I, where y_i represents the pre-defined label following the BIO-tagging schema. Ultimately, by utilizing the annotations in Y, we can obtain the entities of interest.

3.1.2. Multi-modal relation extraction

For a given sentence T and its corresponding image I, which are derived from the same social media post, the goal of NER is to identify the semantic relationship between entities E_{sub} and E_{obj} within the sentence T, and assign them to predefined relation types. Specifically, let $T = (t_1, t_2, \ldots, t_{n-1}, t_n)$ represent an input sentence containing entities E_{sub} and E_{obj}. Our objective is to predict the relationship $y_i \in Y = \{y_1, y_2, \ldots, y_m\}$ between entities E_{sub} and E_{obj} by incorporating the image I. Here, Y represents the set of pre-defined relation labels.

3.2. Semantic feature extraction

For the given sentence T, we employ BERT [44] to acquire word embedding representations of social texts, and subsequently utilize a Fully Connected(FC) layer to map them into a multi-modal semantic space. As a result, we obtain textual representations in the multi-modal space, denoted as $E_t = \{\mu_1, \mu_2, \ldots, \mu_n\} \in \mathbb{R}^{n \times d}$, where n represents the length of the social text and d represents the dimension of the multi-modal semantic space. Furthermore, we utilize self-attention to capture the global features of the text, resulting in the representation $\overline{\mu}$.

For a given social image I, we utilize a visual grounding toolkit [77] to extract its object-level features, and employ a fully connected layer to obtain its feature representation in the multi-modal semantic space, denoted as $E_v = \{v_1, v_2, \ldots, v_{m-1}, v_m\} \in \mathbb{R}^{m \times d}$, where m represents the number of the object feature. Subsequently, we employ self-attention to capture the global features of the images, resulting in the representation $\overline{v}$.

3.3. Modality relevance reasoning

3.3.1. Relevance representation

In this section, we utilize relevance representations based on cross-modal alignment to calculate the relevance weights between modalities. To obtain alignment representations between different modality features, we utilize the following relevance function to compute alignment representations between the features of different modalities, treating them as relevance representations.

$$r(\mu, v, W) = \frac{W\,|\mu - v|^2}{\left\| W\,|\mu - v|^2 \right\|_2} \tag{1}$$

where $W \in \mathbb{R}^{m \times d}$ represents learnable parameters, $|\cdot|^2$ denotes element-wise square operation, and $\|\cdot\|_2$ signifies the l2-norm operation. By utilizing Eq. (1) and the global representations of modalities, we can compute the global correlation representations r_g.

$$r_g = r(\overline{\mu}, \overline{v}, W_g) \tag{2}$$

To simulate human-like discernment of relevance between object features across different modalities, we employ textual-to-visual attention to determine visual object weights that are attentive to textual information.

$$\alpha_{ij} = \mathrm{softmax}\left(\lambda \cdot \frac{[cos(\mu_i, v_j)]_+}{\sqrt{\sum_{j=1}^{L} [cos(\mu_i, v_j)]_+^2}}\right) \tag{3}$$

where λ represents the temperature coefficient, and $[\cdot]_+ = \max(x, 0)$. Next, we can utilize the attended visual features a_j^v for the jth word and the feature representation of the jth word to derive relevance representations based on local semantic alignment.

$$a_j^v = \sum_{i=1}^{K} \alpha_{ij} v_j \tag{4}$$

$$r_l^{i,j} = r(\mu_i, a_j^v, W_l) \tag{5}$$

where $W_l \in \mathbb{R}^{m \times d}$ is a learnable parameter. Compared to r_g, r_l represents a more fine-grained semantic alignment, capturing a better reflection of the degree of correlation between the image and text.

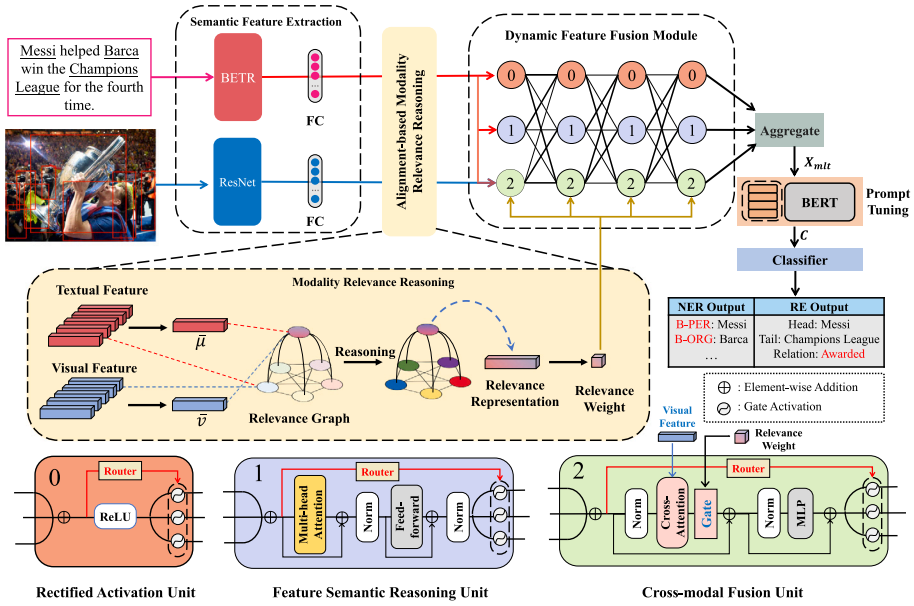


Fig. 3. The overall architecture of our Relevance-Aware Prompt-tuning (RAP) model for multi-modal entity and relation extraction. Please refer to Fig. 5 for the specific details of the Prompt Tuning module.

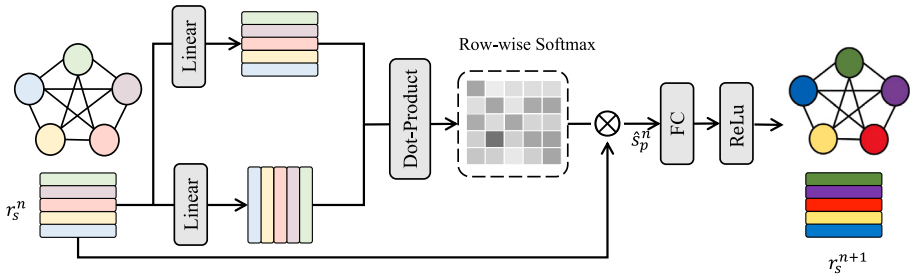


Fig. 4. The proposed Graph reasoning module for computing the modality relevance. We use the local relevance representation r^l and the global relevance representation r^g as initial graph nodes r^0. We compute the edge from node $r_s rc \in r^0$ to $r_t at \in r^0$ by the inner product between incoming and outgoing node representations, followed by a row-wise softmax. After the nth iteration, node r_p^n will be updated to r_p^{n+1} and this process will iterate two times.

3.3.2. Relevance graph reasoning

In general, the local and global correlations between modalities should be consistent, and they model the degree of correlation between modalities at different levels of granularity. Measuring the degree of correlation between different modalities cannot be done unilaterally by using only global or local correlation. Local correlation is calculated based on object-level features, which overlook the scene in the image and the context in the text. Therefore, when evaluating the correlation between two modalities, it is necessary to comprehensively consider both global and local correlation representations. In this paper, we propose a graph-based method to model both global and local correlations and use a graph reasoning algorithm to enable message passing between global and local nodes, thereby taking both correlation representations into account. To model the relations between global and local relevance representations, we construct a relevance graph. The relevance representations are treated as nodes in the graph, and we employ Kuang's [78] method to calculate the edge weights between the nodes, as shown in Fig. 4.

$$e(r_s, r_t, W_{src}, W_{tgt}) = \frac{exp\Big((W_{src}r_s)(W_{tat}r_t)\Big)}{\sum_q exp\Big((W_{src}r_s)(W_{tat}r_t)\Big)} \tag{6}$$

where $W_{src} \in \mathbb{R}^{m\times m}$ and $W_{tgt} \in \mathbb{R}^{m\times m}$ are learnable parameters and r_s and r_t represent two different relevance representations. Subsequently, we propagate the similarity information between both local and global level alignments to update the node representations.

$$\hat{r_s}^n = \sum_s e(r_s^n, r_t^n, W_{src}^n, W_{tgt}^n) \cdot r_t^n \tag{7}$$

$$r_s^{n+1} = \mathrm{ReLU}(W_r^n \, \hat{r_s}^n) \tag{8}$$

where r_s^n represents the feature representation of the node after n iterations of updates. As the number of iterations increases, each node can learn the similarity information from other nodes, and the representations between nodes become more similar.

The modality relevance β can be computed through a FC layer applied to the representation of the global relevance representation r_g^N, as shown in Algorithm. 1.

3.4. Dynamic fusion module

In this section, we introduce three different feature fusion units. They are parallelly deployed at each layer and interconnected using a fully connected manner, forming a routing space. The features of different modalities dynamically select paths in the routing space, adaptively determining the fusion strategy. Each fusion unit consists of a routing gate and different feature fusion operations. Thus, the computation process of each unit can be summarized as follows:

$$\mathcal{H}_i^l = U_i^l(X_i^l) \tag{9}$$

where X_i^l and $\mathcal{H}_i^l$ denote the input and output of the ith$((0 \le i \le 2))$ unit in layer l $(1 \le l \le 3)$. U_i^l represents the ith unit in layer l.

Algorithm 1: Modality Relevance Reasoning(MRR)

Input: Textual feature E_μ and visual Feature E_v
Output: Modality relevance weight β
initialization;
$N \leftarrow n$ // Number of iterations
$iter \leftarrow 0$ // Initialize the iteration counter
/* Compute the global alignment representation */
$g_\mu^l \leftarrow \text{SelfAttention}(E_\mu)$;
$g_v^l \leftarrow \text{SelfAttention}(E_v)$;
Compute r_g according to Eq. (1) ;
/* Compute the local relevance representation */
for $v_i \in E_v$ **do**
  Compute $\alpha_{i,j}$ according to Eq. (3);
  Compute a_j^v according to Eq. (4);
end
for $\mu_i \in E_v$ && $a_j^v \in a^v$ **do**
  Compute local relevance representation $r_l^{i,j}$ according to Eq. (5);
end
$r = r_l \cup r_g$;
/* Update the representation */
while $iter \leq N$ **do**
  for $r_i \in r$ **do**
    for $r_j \in r$ **do**
      Compute the edge weight e_{ij} according to Eq. (6);
    end
    $r_i^{iter} \leftarrow \sum e_{ij} \cdot r_j^{iter}$;
    $r_i^{iter+1} \leftarrow ReLU(W_r r_j^{iter})$;
    $iter \leftarrow iter + 1$;
  end
end
$\beta \leftarrow W_s \cdot r_g^N + b_s$ // Compute the modal relevance

3.4.1. Rectified activation unit (RAU)

When given text with explicit entity information or relation clues, humans can easily and accurately extract entities and relationships from the textual modality. Therefore, it is unnecessary to perform complex fusion representation for these "easy samples". To preserve these discriminative features, we designed a Rectified Activation Unit (RAU) to avoid unnecessary complex encoding for easy samples. The RAU can be seen as a skip connection operation in a neural network.

$$U_0(X) = \text{ReLu}(X) \tag{10}$$

3.4.2. Feature semantic reasoning unit (FSRU)

The main reasons why humans can accurately extract entities and relationships from social media with modality noise are two-fold. The first is that humans possess a wealth of prior knowledge, while the second is their capacity to comprehend and reason with features. Multi-Head Attention (MHA) has been demonstrated to assist models in extracting higher-level semantic information from different attention representation spaces. In this unit, we employ MHA to fuse features from multiple perspectives, followed by a Feed-Forward Network (FFN) for further reasoning.

$$X_{mha} = \text{Norm}(X + \text{MHA}(X)) \tag{11}$$

$$U_1(X) = \text{Norm}(\text{FFN}(X_{mha}) + X_{mha}) \tag{12}$$

3.4.3. Cross-modal fusion unit (CFU)

In this unit, we alleviate the impact of modality noise on the model by utilizing the modality relevance weights β calculated in Section 3.3. We treat the input textual features X as query and the visual cues as key and value, which are input into Cross-Attention (CSA) for cross-modal fusion. The attended visual features X_{av} can be obtained by Eq. (13). It is worth noting that we employ relevance gating to weight X_{csa} for avoiding introducing the irrelevant visual information and modality noise to the model.

$$X_{av} = \text{CMA}\big(\text{Norm}(E_v; X)\big) \tag{13}$$

$$X'_{av} = \text{R}{-}\text{Gate}(\beta) \cdot X_{av} \tag{14}$$

where $\text{R}{-}\text{Gate}(\cdot)$ represents the relevance gate and $\text{R}{-}\text{Gate}(\cdot) = \tanh(\cdot)$. To prevent the gradient vanishing, we apply residual connections and layer normalization and employ Multi-Layer Perceptron (MLP) for further reasoning.

$$X''_{av} = X + X'_{av} \tag{15}$$

$$U_2(X) = X''_{av} + \text{MLP}\Big(\text{Norm}(X''_{av})\Big) \tag{16}$$

3.4.4. Router mechanism

As shown in Fig. 3, we parallelly deploy each unit at every layer, and organize them between layers using fully connected connections, forming a densely connected routing space. Each unit in the routing space has a certain probability of receiving the output signals from all units in the previous layer. This connectivity ensures flexibility and adaptability in feature fusion. Moreover, in the routing space, we employ a soft routing mechanism for forward propagation, which can be viewed as an adaptive path or fusion strategy selection. Specifically, the input to the ith unit in the lth layer is obtained through the following aggregation operation:

$$X_i^l = \begin{cases} X, & l = 0 \\ \sum_{j=0}^{C-1} g_{j,i}^{(l-1)} \mathcal{H}_j^{(l-1)}, & l > 0 \end{cases} \tag{17}$$

Where $X \in \mathbb{R}^{n \times d}$ represents the text modality features inputted into the first layer of the dynamic fusion module, $\mathcal{H}_j^{(l-1)}$ represents the output of the jth cell in the $(l-1)$th layer, and $g_{j,i}^{(l-1)}$ represents the path probability from the jth cell in the $(l-1)$th layer to the ith cell in the lth layer. $g_{j,i}^{(l-1)}$ is calculated by the Router of the jth cell in the $(l-1)$th layer:

$$g_{j,i}^{(l-1)} = \mathcal{R}_i^{(l-1)}(X_i^{(l-1)}) \tag{18}$$

$$\mathcal{R}_i^{(l+1)}(X_i^{(l+1)}) = \text{ReLU}\Big(\tanh\big(\text{MLP}(\frac{1}{M}\sum_{r=1}^{M} X_{i,r}^{(l)})\big)\Big) \tag{19}$$

where $X_{i,r}^{(l)}$ is the rth row vector of X_i^l.

3.5. Path regularization

Intuitively, similar samples should learn similar routing decisions. In other words, when samples have similar semantics or belong to the same category, the paths they learn should be similar. For example, if two samples both are "easy samples" and they have the same relation label, they both do not need to rely heavily on Cross-modal Fusion Unit (CFU) to extract relations between the entities. On the contrary, they would need to increase the path weight of Cross-modal Fusion Unit (CFU). Through this operation, we align the routing distribution with the semantic distribution. To achieve this, we introduce a path regularization term by considering the semantic similarity between samples. The semantic relevance between samples is calculated using the [CLS] token outputs $\hat{\mu}_x$ and $\hat{\mu}_y$ from PLM.

$$L = L_{task} + \gamma \cdot L_{router} \tag{20}$$

$$L_{task} = \begin{cases} -\log\ P(x|y), & \text{task} = \text{NER} \\ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}), & \text{task} = \text{RE} \end{cases} \tag{21}$$

$$L_{router} = \sum_{x,y\in\mathcal{B}} \left(\cos(g_x, g_y) - \cos(\hat{\mu}_x - \hat{\mu}_y)\right)^2 \quad (22)$$

where γ is a hyperparameter, $\mathcal{B}$ denotes a mini-batch and L_{task} represents the loss functions for different downstream tasks. The loss function for the NER task is defined by maximizing the conditional probability of the true sequence y, while the loss function for the RE task is the cross-entropy loss, as shown in Eq. (21). As for the path regularization term (Eq. (22)), we measure the semantic correlation between samples by calculating the cosine similarity between the routing distributions $g_x \in \mathbb{R}^{U^2(L-1)+U}$ and $g_y \in \mathbb{R}^{U^2(L-1)+U}$ and the semantic distributions $\hat{\mu}_x$ and $\hat{\mu}_y$. We then minimize the difference between this cosine similarity and the routing distributions to achieve alignment between the routing and semantic distributions.

3.6. Prompt-tuning and classifier

Although the routing mechanism can minimize the impact of visual noise, irrelevant visual information still significantly affects the model's performance. The reason is that visual noise influences the model's original inductive bias during the backpropagation process, gradually causing the model to deviate from the true distribution, thereby affecting its performance. Prompt tuning is a non-intrusive fine-tuning method, allowing it to fine-tune for specific downstream tasks while maintaining the performance of the backbone model. In this paper, we do not directly use the fused multi-modal features for entity and relation recognition. Instead, we use them as prompts and employ the P-tuning v2 [73] method for entity and relation extraction. However, unlike P-tuning v2, we do not use an embedding layer to obtain prompt vectors. Instead, we use a projection network to map the multi-modal features into prompt vectors. Specifically, after the computation in the L-layer dynamic fusion module, we aggregate the output of each unit in the last layer to obtain a refined multi-modal representation, denoted as $X_{mlt} \in \mathbb{R}^{m\times d}$. Before performing prompt tuning, we use a linear layer to map the length of X_{mlt} to the preset prompt length m_p, resulting in $X'_{mlt} \in \mathbb{R}^{m_p\times d}$. Then, we use two linear layers and a non-linear layer as the projection network. The first linear layer does not change the dimension of X_{mlt}, while the second linear layer expands the last dimension of X_{mlt} to $2 \times NL \times d$, obtaining X_{prompt}, where NL represents the number of layers in BERT. In this paper, we use a 12-layer BERT. We feed the text into BERT and concatenate the output of each Transformer layer with the corresponding prompt vectors from the 12 layers, then input it into the next Transformer layer of BERT, finally obtaining the final output C.

$$X_{prompt} = W_{out} \cdot ReLu(W_{in} \cdot X_{mlt}) \quad (23)$$

$$C = \text{PLM}(X; X_{prompt}) \quad (24)$$

where W_{in} and W_{out} are learnable parameters.

For MNER task, we then feed context representations $C \in \mathbb{R}^{m\times d}$ to a standard CRF layer following [36], which defines the probability of the label sequence y given the input sentence and its associated image:

$$p(y_i|y_0, \ldots, y_{i-1}, C) = \text{CRF}(y_0, \ldots, y_{i-1}, C) \quad (25)$$

For the MRE task, its objective is to extract the relations between entities E_{sub} and E_{obj} from the given input sentence S and its associated image V. Thus, we extract the representations corresponding to the subject entity and object entity from the context representations $C \in \mathbb{R}^{m\times d}$ and utilize softmax for relationship classification.

$$p(y|C) = \text{softmax}([C_s; C_o]) \quad (26)$$

where s, o denotes index of the subject and object entity, respectively.

Table 1
The basic statistic of Twitter-2015 and Twitter-2017.

Entity type	TWITTER-2015			TWITTER-2017		
	train	dev	test	train	dev	test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Miscellaneous	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351

Table 2
The basic statistic of MRE.

Dataset	#Img	#Word	#Sent	#Ent	#Rel
SemEval-2010 Task 8	–	205k	10 717	21 434	9
TACRED	–	1,823k	53 791	152 527	41
MNRE	10 089	172k	14 796	20 178	31

4. Experiments

4.1. Dataset

The Twitter-2015 dataset [17], Twitter-2017 dataset [21], and MRE dataset [25] are derived from social texts on Twitter, and they have been extensively utilized in recent years to assess the performance of models in MNER and MRE tasks. The statistical and experimental details of the datasets are illustrated in Tables 1 and 2.

4.2. Settings

In this chapter, we present the training configurations and detailed settings for different datasets. We utilize the BERT-base-uncased model as the text encoder and employ ResNet101 as the image encoder. The model parameters are optimized using the AdamW optimizer, with the learning rate linearly warmed up to its maximum value during the first 10% of gradient updates. We set the number of image objects, denoted as m to 3. For the MNER task, we set the batch size to 8 and train the model for 30 epochs with a learning rate of 3e-5. For the RE task, we set the batch size to 16 and train the model for 15 epochs using a learning rate of 2e-5. Finally, we select the best model based on performance on the validation set and evaluate it on the test set.

4.3. Compared methods

To demonstrate the effectiveness of our RAP, we compared it with three groups of baselines: text-based baselines, vanilla multimodal baselines, and multi-modal baselines considering modality noise.

- Text-based baselines: CNN-BiLSTM-CRF [79], HBiLSTM-CRF [36], BERT [44] and BERT-CRF for MNER; PCNN [80] and MTB [5] for MRE;
- Vanilla multimodal baselines: AdapCoAtt [17], MAF [81], FMIT [19] and UMT [58], BGA-MNER [82] for MNER; BERT+SG [26], MEGA [26] and VisualBERT [83] for MRE;
- Multi-modal baselines considering modality noise: UMGF [64], HVPNet [24], GEI [57] and MKGformer [82].
- LLM-based baselines: CoT [84], UMIE [85], MoRe [63]

4.4. Results

Table 3 presents the overall results of our model on the test sets of the three different datasets. We conducted experiments on MEGA, HVPNet, and RAP in the same experimental environment. The performance of other baselines is referenced from the results reported in their papers. From Table 3, we can draw the following conclusions:

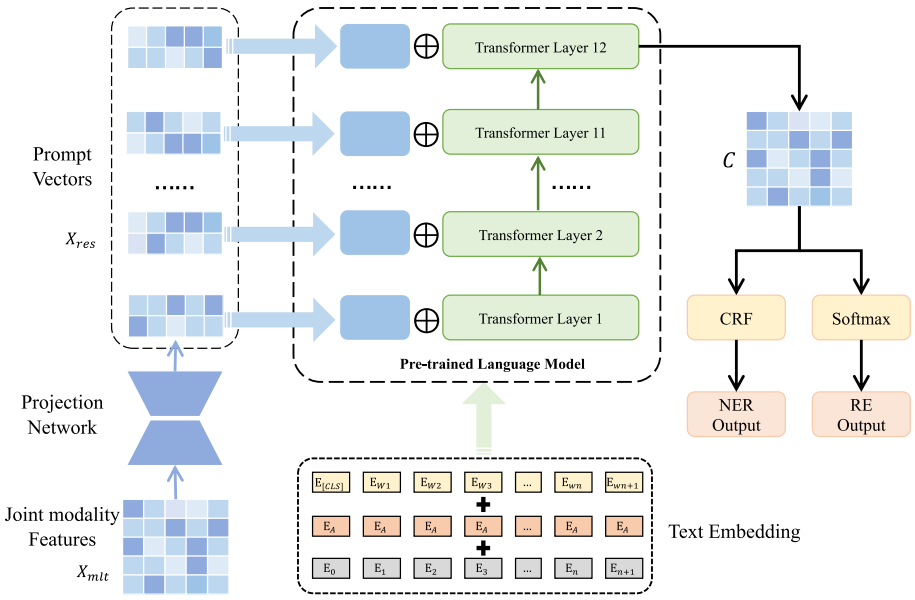


Fig. 5. The architecture of the downstream classifier. We will fuse the obtained joint multimodal features using a projection network to generate 12 prompt vectors, and concatenate them with the hidden states of each Transformer layer in BERT.

Table 3
Performance comparison of different competitive baseline approaches for NER and RE. Since the original results of UMT, UMGF and MEGA only involve single extraction task, we reproduce their public code for more comprehensive comparison.

Modality	Methods	Twitter-2015			Twitter-2017			MRE		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Text	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37	–	–	–
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37	–	–	–
	BERT	68.30	74.61	71.32	82.19	83.72	82.95	75.39	61.17	57.03
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44	–	–	–
	PCNN	–	–	–	–	–	–	62.85	49.69	55.49
	MTB	–	–	–	–	–	–	64.46	57.81	60.96
Text+Image	AdapCoAtt	72.75	68.74	70.69	84.16	80.24	82.15	–	–	–
	UMT	71.67	75.23	73.41	85.28	85.34	85.31	62.93	63.88	63.46
	BERT+SG	–	–	–	–	–	–	65.95	62.65	62.80
	MEGA	71.87	72.95	72.41	84.33	84.62	84.47	64.77	68.39	66.53
	VisualBERT	68.84	71.39	70.09	84.06	85.39	84.72	57.15	59.48	58.30
	MAF	71.86	75.10	73.42	86.13	86.38	86.25	–	–	–
	FMIT	74.18	75.03	74.60	85.55	85.29	85.42	–	–	–
	BGA-MNER	78.60	74.16	76.31	**87.71**	87.71	87.71	–	–	–
	UMGF	**74.49**	75.21	74.85	86.54	84.50	85.51	64.38	66.23	65.29
	HVPNet	73.91	76.82	75.34	85.72	87.12	86.41	79.37	79.82	79.59
	GEI	73.39	75.51	74.43	87.50	86.01	86.75	–	–	–
	MKGformer	–	–	–	–	–	–	82.67	81.25	81.95
	CoT(ChatGPT)	–	–	76.53	–	–	87.79	–	–	66.42
	UMIE(Flan-T5)	–	–	76.53	–	–	87.79	–	–	66.42
	MoRe(CLIP+MoE)	–	–	**79.21**	–	–	**90.67**	–	–	68.60
	RAP(ours)	74.38	**78.92**	76.58	87.63	**87.89**	87.76	**83.59**	**82.67**	**83.13**

(1) Most multimodal methods have achieved better performance than methods based solely on textual modality. This indirectly demonstrates that visual modality information can assist textual modality in entity and relation extraction. Moreover, the improvement brought by visual modality information is more pronounced in the RE task. Multimodal relation extraction methods have achieved a performance improvement of 2.8% to 22.17% in F1-score compared to state-of-the-art methods based solely on textual modality. Although AdapCoAtt and VisualBERT did not perform well and even fell short of some single-modality methods, AdapCoAtt is one of the earliest MNER methods, which overlooked semantic alignment between different modalities and did not consider issues such as modality noise. Additionally, VisualBERT is a general-purpose visual language model, not specifically designed for MNER and MRE. Furthermore, VisualBERT is pretrained on four multimodal tasks, including Visual Question Answering (VQA) and Visual Commonsense Reasoning (VCR), making its direct use in MNER and MRE tasks reasonable. Therefore, multimodal-based entity and relation extraction methods can be better applied to social media datasets.

(2) The issue of modality noise limits the performance of MNER and MRE. While multimodal entity and relation extraction methods have achieved decent performance, we have found that multimodal methods that consider modality noise have achieved more significant improvements. Most multimodal approaches have overlooked the issue of noise propagation caused by irrelevant images, and methods considering modality relevance tend to outperform other multimodal MNER and MRE methods. For example, UMFG introduces visual regions features corresponding to text entities rather than the features from the entire image to mitigate the impact of visual modality noise

on the model, and it has outperformed vanilla multimodal baselines. Besides, HVPNet treats hierarchical visual prefixes as prompts to reduce sensitivity to visual noise, while MKG introduces a relevance-aware fusion module to alleviate the problem of modality noise. Most of them achieved a 16.56% and 16.66% improvement in F1 score compared to UMGF in the MRE task. Notably, BGA-MNER achieved results similar to our method on the MNER task, despite not accounting for modal noise. The main reason for this is that BGA-MNER not only uses ViT-B as the visual encoder, but also stacks 11 layers of BGA modules and Transformer layers, and incorporates two auxiliary tasks, text-to-image and image-to-text, during training. This allows it to have better visual features and a larger number of parameters, making its strong performance understandable. However, our method, with simpler training objectives and fewer model parameters, still outperforms it. Our method does not rely on simple dot-product attention to calculate relevance weights, nor does it solely introduce relevant visual features to avoid modality noise. Instead, we compute the relevance representations through global and local semantic alignment and construct relevance graphs to compute modality relevance. Finally, we employ dynamic routing mechanisms and prompt learning for multimodal entity and relation extraction.

(3) The relevance reasoning and dynamic routing mechanism are effective strategies for mitigating modal noise. While existing baselines that consider modality noise have achieved satisfactory results, RAP is the first method to introduce visual information based on the complexity of the samples, which makes our approach more flexible and minimizes the introduction of modal noise as much as possible. Although UMGF introduces image segments relevant to entities, and GEI uses a heterogeneous graph interaction network to capture visually relevant information, which can alleviate modal noise to some extent, they inevitably introduce noise when the image itself is entirely unrelated. In contrast, our dynamic routing mechanism can selectively introduce visual information when the sample requires visual assistance through path selection. If the image is entirely irrelevant, the relevance reasoning module outputs a very small value to prevent visual information from misleading the model. Furthermore, we do not directly use the computed multimodal features for entity and relation extraction. Instead, we input them as prompts into PLMs and perform extraction using prompt-based learning, which further helps mitigate the negative impact of modality noise on the model.

(4) Our model is the only method that achieves stable SOTA performance in both MNER and MRE tasks simultaneously. Objectively, we observed a more significant improvement in the MRE task, while the performance in the MNER task only saw an increase of 1.01% and 1.24% in F1 score compared to the aforementioned baselines. However, comparing directly to these baselines is unfair. Many multimodal named entity recognition methods are tailored for specific tasks, and they do not need to consider performance on the MRE task. Therefore, they can focus solely on the characteristics of the MNER task for careful design. In contrast, our method employs a multitask architecture that is applicable to both MNER and MRE tasks simultaneously. Although HVPNet is also a multitask architecture, when conducting 10 repeated experiments in the same experimental environment, we found that HVPNet's performance on the MRE task still lags behind some MRE methods. However, it is worth noting that HVPNet still achieves state-of-the-art performance on the MNER task. Therefore, when considering unified extraction capability, our method still holds a significant advantage.

(5) Even compared to methods based on large language models (LLMs), our approach still holds certain advantages. Although there is some gap between our method and LLM-based methods on the MNER task, especially MoRe, it is inherently unfair to directly compare them. LLM-based methods have larger model scales and are generally pre-trained on larger corpora, making their superior performance on generative extraction tasks understandable. Despite this, our method still achieves performance comparable to CoT and UMIE on both the MNER and MRE tasks, which demonstrates its significant advantages in my view. MoRe's F1 score is about 3% higher than ours on the two MNER datasets, primarily because it uses additional data and knowledge during training. In contrast, our method does not incorporate extra data or use techniques like MoE to boost performance, instead relying more on the dynamic fusion network's ability to handle modal noise to improve upon existing work. However, although the aforementioned methods achieve success on the MNER task through larger model sizes and additional data, their performance on the MRE task is even lower than that of MKGformer, let alone our method. Therefore, in terms of unified extraction capability, our method is lighter, more generalizable, and more robust.

5. Analysis

5.1. Ablation experiment

In this section, we conduct ablation experiments to further explore the effectiveness and influential factors of the proposed module.

5.1.1. Effect of the number of routing layers

To investigate the optimal number of Routing Layers in the dynamic fusion module, we conducted ablation experiments by varying the number of layers $L \in \{2, 3, 4, 5, 6\}$ as shown in Table 4. We observed that increasing the number of layers can improve the performance of our model until the number of layers exceeds 3 or 4. This can be attributed to the fact that increasing the number of layers provides a broader path space, increasing the possibility of exploring higher-level feature representations. When $L > 4$, we observe that the model's performance saturates on all three datasets. As we further increase the number of layers, their F1 scores remain relatively stable. However, when $L > 4$, the model's performance on the Twitter-2014 and MRE datasets actually declines. This could be attributed to the fact that as the number of layers increases, the differences between different path selections diminish. The advantage of the routing mechanism cannot be fully manifested, and makes it challenging for the model to distinguish between simple and complex samples. Increasing the number of layers to six or more in our model will likely result in the input of features into each of the calculation units we have designed. The core principle of our proposed method is to minimize the involvement of visual modality features in the feature calculation, as much as possible, while maintaining a complete context. This is achieved by carefully selecting the path through which features are processed. A straightforward approach to achieve this is by assigning a minimal weight to the dynamic fusion module, thereby ensuring that feature computation predominantly relies on the other two calculation units. However, as the number of layers increases, this approach becomes less effective, diminishing the overall performance of our method. Furthermore, if the number of layers continues to grow, issues such as feature distortion and gradient vanishing may arise, leading to a degradation in model performance. This highlights the importance of balancing the depth of the network to avoid these challenges. Based on our experimental results, we find that the optimal performance for the MNER task is achieved when $L = 4$, while for the MRE task, the best performance occurs when $L = 3$.

5.1.2. Effect of the different modules and units

To investigate the impact of the Modality Relevance Reasoning module and each unit in the Dynamic Fusion Module on experimental performance, we conducted several ablation experiments as shown in Table 5. Among them, "MRR" represents the model performance when we remove the Modality Relevance Reasoning module from the complete model (ALL) and set the value of the Relevance Gate to 1. "RAU", "FSRU", and "CFU" represent the experimental results when removing three different computational units from the complete model.

As shown in Table 5, when we remove the MRR module, our model no longer considers the relevance between modalities, resulting in

Table 4
Ablation on the number of the routing layer L.

Layer number	Twitter-2015			Twitter-2017			MRE		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
2	74.35	78.62	76.42	87.42	87.79	87.60	83.49	82.47	82.97
3	74.37	78.67	76.46	87.55	87.84	87.69	83.59	**82.67**	**83.13**
4	74.38	**78.92**	**76.58**	87.63	**87.89**	**87.76**	83.57	82.23	82.89
5	74.37	78.82	76.53	**87.65**	87.86	87.75	**83.60**	81.55	82.56
6	**74.40**	78.79	76.53	87.62	87.84	87.73	83.55	81.55	82.54

Table 5
Ablation on different module and Units."MRR" stands for removing the Modality Relevance Reasoning module; "RAU", "FSRU" and "CFU" stands for removing the three unit we proposed in Section 3.4.

Methods	MRR	RAU	FSRU	CFU	ALL
Twitter-2015	74.56(−2.02)	75.66(−0.92)	75.15(−1.43)	72.47(−4.11)	76.58
Twitter-2017	83.83(−3.93)	86.68(−1.08)	85.23(−2.53)	84.55(−2.21)	87.76
MRE	81.45(−1.68)	81.97(−1.16)	78.81(−4.32)	75.62(−7.51)	83.13

Table 6
Ablation on the value of the γ.

γ	0	0.5	1	1.5	2	2.5	3
Twiter2014	75.98	76.27	76.45	**76.58**	76.52	76.44	76.48
Twitter2015	87.45	87.78	**87.81**	87.76	87.79	87.58	87.48
MRE	82.71	82.68	82.74	**83.13**	82.87	82.89	82.63

visual and textual features having equal weights. If the images and text are unrelated, the model becomes susceptible to the influence of modality noise. At this point, our model's performance decreases by approximately 2% to 4% in F1 score. On the other hand, when we remove any one of the RAU, FSRU, and CFU computational units, the performance of the model significantly declines. We believe that this observation is reasonable because, when the number of layers in the network remains constant, reducing one of the computing units alone diminishes the dimensionality of the path space, which consequently reduces the diversity of possible path selections. Furthermore, our findings indicate that the performance of the model is more significantly impacted by the computing units in FSRU and CFU. We hypothesize that this is due to the multi-layer MLP structure inherent in these two units, which enhances the network's nonlinear capabilities. In contrast, RAU lacks this structural feature, although it helps alleviate the vanishing gradient problem, contributing to the overall stability of model performance. It is noteworthy that the FSRU and CFU units exert a particularly strong influence on the MRE task, with CFU having a more substantial effect. We suggest that this is primarily because the MRE task relies heavily on the interaction between different modalities. The computing units in FSRU and CFU are essential for enabling such interactions. Without a complete contextual understanding, the model struggles to accurately predict relationships between entities. If information from the visual modality fails to provide sufficient context, the model's likelihood of making errors increases. In contrast, the NER task is less dependent on contextual information, and under similar conditions, the model's performance is less sensitive to changes in the computing units.

5.1.3. Effect of the balance parameters γ

In the routing mechanism proposed in this paper, we introduce path regularization to ensure that similar samples follow similar paths during learning. To investigate the impact of different balance parameters, denoted as γ, on model performance, we set γ to various values, as shown in Table 6, and analyze its effects on the MNER and MRE tasks. Our experiments show that as γ increases, the model's performance steadily improves. This trend suggests that path regularization plays a crucial role in helping the model make better path decisions across different samples. When γ approaches zero, however, the model lacks a supervisory signal to guide dynamic routing, leaving it entirely dependent on the neural network's learning ability. This reliance leads to unstable model performance, which is highly sensitive to noise. Moreover, with γ near zero, the model's optimization objective focuses solely on classification accuracy, neglecting the development of an effective path strategy to handle visual noise. As a result, the final F1 score deteriorates. Conversely, when γ exceeds a value of 2, model performance begins to decline. This suggests that excessive path regularization forces similar samples to adopt identical paths, disregarding subtle distinctions between closely related samples. Such an approach hampers the model's ability to capture fine-grained semantic differences and imposes an unrealistic constraint, compelling similar samples to share the same fusion path. Additionally, we observe that the MRE task is more sensitive to path regularization than the MNER task. As indicated in Table 6, the performance of the MRE task exhibits more pronounced fluctuations with varying γ values compared to the MNER task. This phenomenon can be explained through the concept of distant supervision: in relation extraction (RE) tasks, samples with the same entity pairs typically share the same relation labels. Similarly, in the MRE task, samples with similar textual and visual semantics are more likely to have the same relation labels. Path regularization enables the model to learn similar paths, grouping these samples into closer categories, thereby introducing prior knowledge that helps the model make more informed decisions.

5.2. Path analysis

Intuitively, samples that are similar in semantics and images should choose the same paths. Therefore, we employed path regularization in Section 3.5, which not only facilitates better learning of path selection by the model but also ensures the consistency between semantic distribution and routing distribution. To verify whether our model can effectively learn this consistency, we concatenate the routing vectors g from each layer of every test sample to obtain the current sample's path representation. Then, we utilize the t-SNE algorithm to reduce the dimensionality of the path representations of each test sample and print them on a two-dimensional plane, as shown in Fig. 6.

Although we did not display the text of the samples on Fig. 6, we found that in social media posts, when the images are similar, the overall semantics of the text are also similar. However, the reverse may not necessarily hold true. We can observe from Fig. 6 that samples with similar images have closer distances between them, indicating that they have learned similar paths. Additionally, their relation targets are also similar, such as /per/loc/place_of_residence, /org/loc/locate_at, and /loc/loc/contain. This indirectly demonstrates that their textual semantics are also similar. Specifically, in the green dashed box, the images all contain athletes wearing jerseys. Generally, these samples are likely describing relationship types between players and teams. In contrast, the backgrounds in the yellow dashed box are similar to

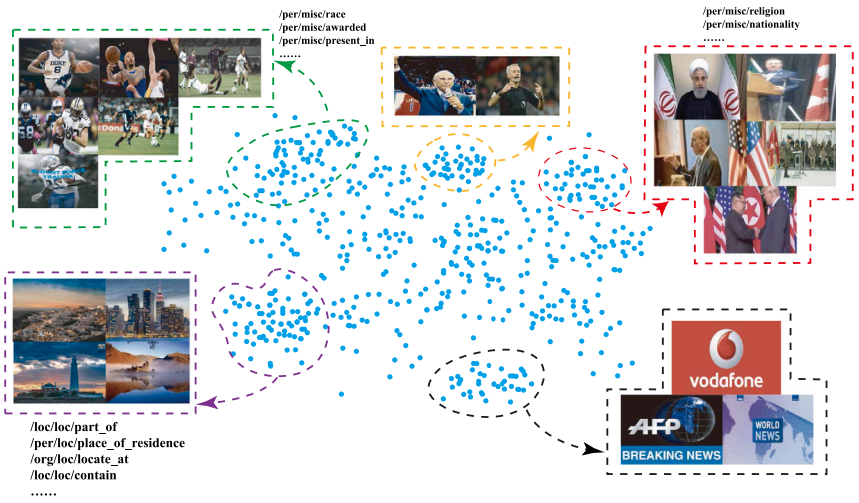


Fig. 6. The visualization of the gate vector in the dynamic fusion module.

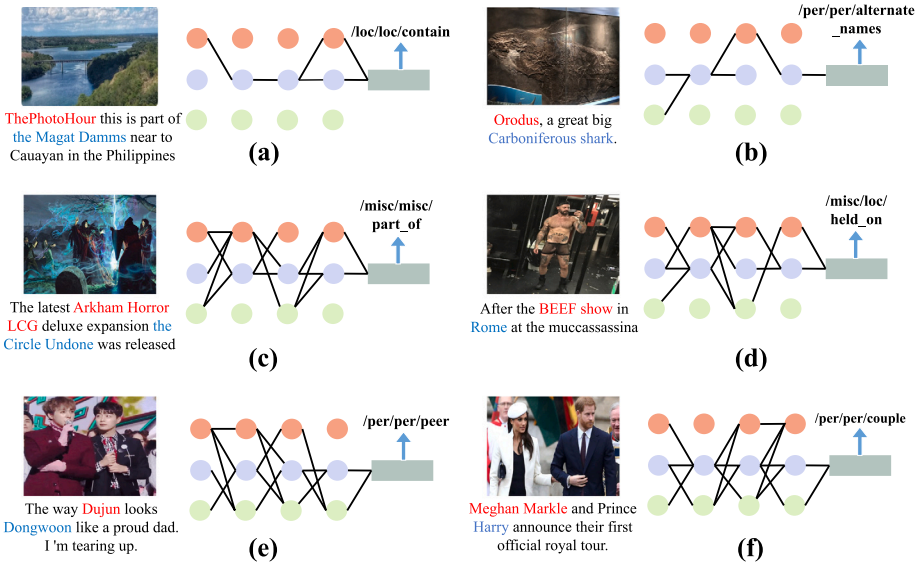


Fig. 7. Visualize the path of the router.

those in the green dashed box, but these images do not feature jerseys, footballs, or basketballs. Therefore, they do not belong to the same cluster as the samples in the green dashed box. However, due to the semantic similarity in the text, their positions are very close to the samples in the green dashed box. It is very clear that the images in the red dashed box all feature flags and people. Therefore, most of these samples likely describe relationships between people and countries, such as /per/misc/nationality, etc. Therefore, our model has learned similar paths for similar samples.

5.3. Qualitative analysis

In order to address the issue of modality noise, we employed a dynamic routing mechanism to enable the model to learn different paths for different samples. For easy samples, our model is designed to rely as much as possible on the textual modality for entity and relation extraction, which helps minimize the introduction of modality noise. To further demonstrate the effectiveness of the dynamic fusion strategy, we visualize the paths with significant weights in the MRE task, as illustrated in Fig. 7.

Fig. 7(a) and Fig. 7(b) represent similar simple samples. Specifically, in Fig. 7(a), there is a very obvious trigger word “part of” in the text, and both entities are of type “Location”. This allows the model to accurately extract the relationship between entities as /loc/loc/contain based solely on textual features. At this point, the paths selected by dynamic routing are quite consistent, and they mainly rely on RAU and FSRU. Because they do not require assistance from visual features, this also reduces the involvement of CFU.

Apart from Fig. 7(a) and Fig. 7(b), the samples belong to hard samples. However, the images and text in Fig. 7(c) and Fig. 7(d) are weakly correlated, rendering the image becomes modality noise. In this scenario, we need to rely more on FSRU to learn the semantics of sentences and reduce the involvement of CFU to avoid introducing modality noise. Although the paths selected by the model still pass through CFU, its involvement is limited, and we can observe more FSRU participating in feature calculation.

For Fig. 7(e) and Fig. 7(f), they require the combination of visual features to accurately extract the entity relations. In Fig. 7(e), the text contains the word "dad”, which makes it very easy for the model to incorrectly classify this sample as "/per/per/parents”. At this point, visual features are crucial. If the model could recognize that the two visual objects are more similar to peers rather than adults and children, it might be able to correct this mistake. Similarly, in Fig. 7(b), it is not possible to determine the entity types solely based on the text, so

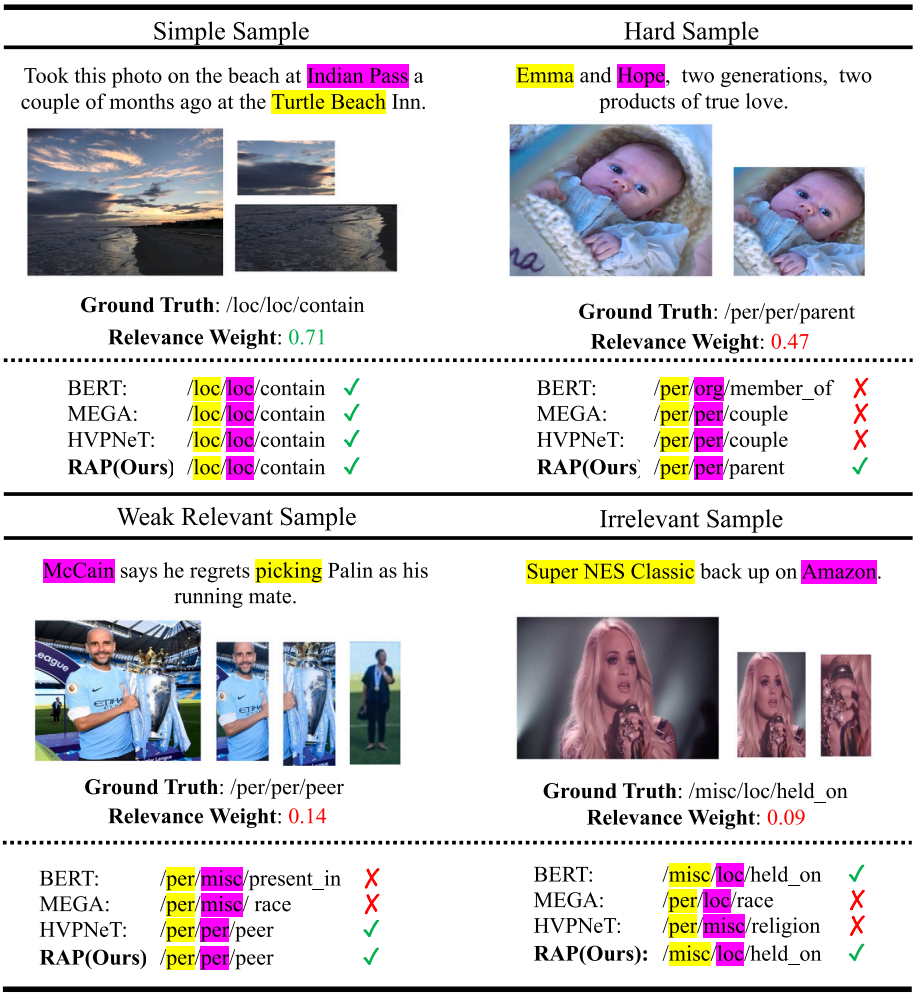


Fig. 8. The case study on the MRE dataset.

it also requires more assistance from visual features. As expected, the paths selected by the model for these two samples also pass through CFU more frequently.

In summary, the dynamic routing mechanism has performed as expected. It selects appropriate paths for feature calculation for different types of samples, allowing our model to better handle the impact of mismatched image-text pairs.

5.4. Case study

In this paper, we selected four representative samples for quality analysis in MRE experiments to illustrate the superiority of our method. As shown in Fig. 8, for the easy sample, we can straightforwardly infer that the "Turtle Beach" is within "Indian Pass" and, therefore, there exists a "contain" relationship, solely through the social text. However, for challenging samples, the text modality often lacks complete context and may not even form a complete sentence. Without considering the visual modality information (the baby in the image), the model might mistakenly categorize 'Hope' as an organization, which could lead the model to incorrectly assume that "Emma" is a member of "Hope". Since MEGA and HVPNet cannot effectively achieve semantic alignment, they only consider the words "generations" and "love" in the sentence, predicting a "couple" relationship between them.

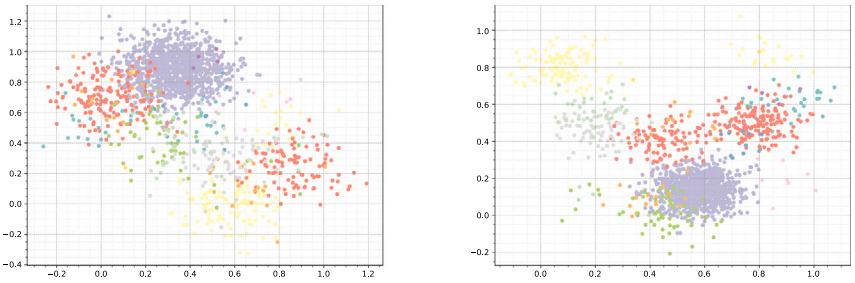
On the other hand, when there exists relatively weak relevance between image-text pairs, the image can still provide partially useful information for the text modality. As shown in Fig. 8, although the relevance weight for the weak relevant sample are low, both people in the image are recognized. HVPNet and RAP can achieve semantic alignment using these two object-level visual features, informing the model that both entities belong to the "person" entity type rather than "organization".

Irrelevant samples make up the majority of the training and testing data, with lower correlation weights, and the visual information actually introduces modality noise, severely affecting HVPNet and MEGA's performance. We address this issue by computing relevance weights and applying weighted processing to the visual information, thereby minimizing the impact of such modality noise.

5.5. Visualizational analysis

To investigate whether P-tuning v2 can effectively handle modality noise, we conducted a visual analysis of RAP in MRE experiments. We directly used the joint modality features X_{mlt} calculated through dynamic routing for relationship extraction without mapping it to prompt vectors X_{prompt}. Finally, we obtained the output C from the last layer of BERT. We performed t-SNE dimension reduction separately for the C obtained in these two different experimental setups and visualized them on a 2D plane. It is worth noting that we did not visualize each row vector of C, but rather, we found the indices corresponding to the entities and concatenated the vectors corresponding to these two indices on C as a data point.

As shown in Fig. 9, we visualized the test sets for two different experimental setups. We found that, while Fig. 9(a) can distinguish different clusters, the boundaries between the clusters are not very



(a) RAP(without P-tuning v2) (b) RAP(all)

Fig. 9. Visual analysis of the impact of P-tuning v2 on RAP.

distinct, and the points within each cluster are relatively far from the cluster center. On the other hand, P-tuning v2 makes the boundaries between clusters in C clearer, and the points within each cluster are closer. This is one of the reasons P-tuning v2 achieves better results. In conclusion, we believe that P-tuning v2 further mitigates the impact of modality noise on the model in our approach.

6. Limitations

It is evident that, compared to traditional methods, our approach can more flexibly handle modal noise issues in multimodal information extraction tasks, thanks to the routing mechanism that dynamically selects computational units. When visual information is needed for assistance, the model will increase the weight of the cross-modal fusion units, allowing visual information to provide more clues. Moreover, our method is easy to couple with different language models, especially those that can be optimized through prompt tuning. However, our method still has its limitations. First, the dynamic fusion module requires a predefined number of network layers. The optimal number of layers varies across different datasets and task scenarios, thus necessitating multiple experiments and manual adjustments based on specific tasks. Second, our method currently only supports language models based on the Transformer encoder, hence it has poor adaptability for large language models, which is also the main reason why we do not use large models as the backbone in our experiments. Lastly, due to the lack of a large amount of multimodal information extraction data for pre-training, our model is far behind large models in scale, and therefore, cannot be compared in terms of performance. However, among models of the same scale, our method has significant advantages.

Although our method addresses the issue of modal noise to some extent, it still faces limitations when the short text lacks sufficient context. In such cases, performance on Named Entity Recognition (NER) and Machine Reading Comprehension (MRC) tasks can be hindered. There are two primary reasons why humans are able to comprehend these short texts despite their lack of context. First, humans possess a vast repository of common sense, which enables them to fill in missing information across modalities and establish connections between them. Numerous studies have explored the integration of external knowledge to improve model performance, with promising results. Second, humans exhibit advanced visual reasoning and imaginative capabilities. For instance, when presented with an image of two individuals of similar age but different genders holding hands, humans would likely infer that they are a couple. These abilities, however, remain a bottleneck for lightweight language models. While multimodal large language models have the potential to address this issue to some degree, a fully mature solution is still lacking.

7. Conclusions

In this paper, we proposed the Relevance-Aware Prompt-tuning (RAP) method for MNER and MRE, aiming to extract valuable entity and relationship information from noisy social media data. Handling modality noise and effectively leveraging valuable visual cues is currently the primary challenge in MNER and MRE tasks. The dynamic routing mechanism adaptively selects the appropriate feature fusion path in the routing space based on the modality relevance between image-text pairs. This allows the model to rely as much as possible on text to determine entity and relation types, minimizing the involvement of visual features, which indirectly prevents visual features from mismatched images. We employ relevance graph inference to more accurately calculate modality relevance and utilize this relevance coefficient to weight the visual modality during multimodal fusion, which helps mitigate the impact of mismatched images on the model. To better eliminate modality noise, the obtained joint modality features will be mapped to prompt vectors and employed using the P-tuning approach for MNER and MRE tasks. Experimental results demonstrate the satisfactory performance of our method on MNER and MRE tasks.

CRediT authorship contribution statement

Zhenbin Chen: Writing – original draft, Writing – review & editing, Visualization, Project administration, Methodology, Investigation, Conceptualization. **Zhixin Li:** Supervision, Writing – review & editing, Project administration, Methodology, Conceptualization. **Mingqi Liu:** Software, Visualization. **Canlong Zhang:** Writing – review & editing. **Huifang Ma:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Nos. 62276073, 61966004, 62266009), Guangxi Natural Science Foundation, China (No. 2019GXNSFDA245018), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, China, Innovation Project of Guangxi Graduate Education, China (No. YCBZ2023055), and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, China.

Data availability

The authors do not have permission to share data.

References

[1] N. Peng, M. Dredze, Improving named entity recognition for chinese social media with word segmentation representation learning, 2016, arXiv preprint arXiv:1603.00786.

[2] S. Jia, L. Ding, X. Chen, Y. Xiang, et al., Incorporating uncertain segmentation information into Chinese NER for social media text, 2020, arXiv preprint arXiv:2004.06384.

[3] C. Jia, Y. Zhang, Multi-cell compositional LSTM for NER domain adaptation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5906–5917.

[4] W. Tang, B. Xu, Y. Zhao, Z. Mao, Y. Liu, Y. Liao, H. Xie, UniRel: Unified representation and interaction for joint relational triple extraction, 2022, arXiv preprint arXiv:2211.09039.

[5] L.B. Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, 2019, arXiv preprint arXiv:1906.03158.

[6] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77.

[7] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, LUKE: Deep contextualized entity representations with entity-aware self-attention, 2020, arXiv preprint arXiv:2010.01057.

[8] E. Bassignana, B. Plank, What do you mean by relation extraction? A survey on datasets and study on scientific relation classification, 2022, arXiv preprint arXiv:2204.13516.

[9] Z. Li, Y. Sun, J. Zhu, S. Tang, C. Zhang, H. Ma, Improve relation extraction with dual attention-guided graph convolutional networks, Neural Comput. Appl. 33 (2021) 1773–1784.

[10] D. Qiu, Y. Zhang, X. Feng, X. Liao, W. Jiang, Y. Lyu, K. Liu, J. Zhao, Machine reading comprehension using structural knowledge graph-aware network, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5896–5901.

[11] Y. Zhao, J. Zhang, Y. Zhou, C. Zong, Knowledge graphs enhanced neural machine translation, in: Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4039–4045.

[12] L. Huang, L. Wu, L. Wang, Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward, 2020, arXiv preprint arXiv:2005.01159.

[13] Z. Li, Z. Peng, S. Tang, C. Zhang, H. Ma, Text summarization method based on double attention pointer network, IEEE Access 8 (2020) 11279–11288.

[14] T. Xian, Z. Li, Z. Tang, H. Ma, Adaptive path selection for dynamic image captioning, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 5762–5775.

[15] T. Xian, Z. Li, C. Zhang, H. Ma, Dual global enhanced transformer for image captioning, Neural Netw. 148 (2022) 129–141.

[16] X. Xie, Z. Li, Z. Tang, D. Yao, H. Ma, Unifying knowledge iterative dissemination and relational reconstruction network for image–text matching, Inf. Process. Manage. 60 (1) (2023) 103154.

[17] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018, pp. 5674–5681.

[18] O. Arshad, I. Gallo, S. Nawaz, A. Calefati, Aiding intra-text representations with visual context for multimodal named entity recognition, in: Proceedings of the 2019 International Conference on Document Analysis and Recognition, IEEE, 2019, pp. 337–342.

[19] J. Lu, D. Zhang, P. Zhang, Flat multi-modal interaction transformer for named entity recognition, 2022, arXiv preprint arXiv:2208.11039.

[20] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, Q. Li, Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1038–1046.

[21] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, Visual attention model for name tagging in multimodal social media, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1990–1999.

[22] S. Cui, J. Cao, X. Cong, J. Sheng, Q. Li, T. Liu, J. Shi, Enhancing multimodal entity and relation extraction with variational information bottleneck, 2023, arXiv preprint arXiv:2304.02328.

[23] H. Wan, M. Zhang, J. Du, Z. Huang, Y. Yang, J.Z. Pan, FL-MSRE: A few-shot learning based approach to multimodal social relation extraction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 13916–13923.

[24] X. Chen, N. Zhang, L. Li, Y. Yao, S. Deng, C. Tan, F. Huang, L. Si, H. Chen, Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction, 2022, arXiv preprint arXiv:2205.03521.

[25] C. Zheng, Z. Wu, J. Feng, Z. Fu, Y. Cai, Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts, in: Proceedings of the 2021 IEEE International Conference on Multimedia and Expo, IEEE, 2021, pp. 1–6.

[26] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, T. Wang, Multimodal relation extraction with efficient graph alignment, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5298–5306.

[27] Q. Zhao, T. Gao, N. Guo, TSVFN: Two-stage visual fusion network for multimodal relation extraction, Inf. Process. Manage. 60 (3) (2023) 103264.

[28] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investig. 30 (1) (2007) 3–26.

[29] R. Malouf, Markov models for language-independent named entity recognition, in: Proceedings of the 6th Conference on Natural Language Learning 2002, 2002, pp. 1–4.

[30] N. Kambhatla, Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, 2004, pp. 178–181.

[31] F.M. Suchanek, G. Ifrim, G. Weikum, Combining linguistic and statistical analysis to extract relations from web documents, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 712–717.

[32] R. Bunescu, R. Mooney, A shortest path dependency kernel for relation extraction, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 724–731.

[33] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[34] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[35] S. Zhang, D. Zheng, X. Hu, M. Yang, Bidirectional long short-term memory networks for relation classification, in: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015, pp. 73–78.

[36] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, 2016, arXiv preprint arXiv:1603.01360.

[37] A. Bharadwaj, D.R. Mortensen, C. Dyer, J.G. Carbonell, Phonologically aware neural model for named entity recognition in low resource transfer settings, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1462–1472.

[38] V. Yadav, R. Sharp, S. Bethard, Deep affix features improve neural named entity recognizers, in: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics, 2018, pp. 167–172.

[39] M. Miwa, M. Bansal, End-to-end relation extraction using lstms on sequences and tree structures, 2016, arXiv preprint arXiv:1601.00770.

[40] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.

[41] H. She, B. Wu, B. Wang, R. Chi, Distant supervision for relation extraction with hierarchical attention and entity descriptions, in: Proceedings of the 2018 International Joint Conference on Neural Networks, IEEE, 2018, pp. 1–8.

[42] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, Z. Jin, Classifying relations via long short term memory networks along shortest dependency paths, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1785–1794.

[43] J. Yu, B. Bohnet, M. Poesio, Named entity recognition as dependency parsing, 2020, arXiv preprint arXiv:2005.07150.

[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[45] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, J. Li, A unified MRC framework for named entity recognition, 2019, arXiv preprint arXiv:1910.11476.

[46] M.E. Peters, M. Neumann, R.L. Logan IV, R. Schwartz, V. Joshi, S. Singh, N.A. Smith, Knowledge enhanced contextual word representations, 2019, arXiv preprint arXiv:1909.04164.

[47] J. Li, Y. Katsis, T. Baldwin, H.-C. Kim, A. Bartko, J. McAuley, C.-N. Hsu, SPOT: Knowledge-enhanced language representations for information extraction, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 1124–1134.

[48] B. Li, D. Yu, W. Ye, J. Zhang, S. Zhang, Sequence generation with label augmentation for relation extraction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 13043–13050.

[49] C. Zheng, Z. Wu, T. Wang, Y. Cai, Q. Li, Object-aware multimodal named entity recognition in social media posts with adversarial learning, IEEE Trans. Multimed. 23 (2020) 2520–2532.

[50] F. Xu, L. Zeng, Q. Huang, K. Yan, M. Wang, V.S. Sheng, Hierarchical graph attention networks for multi-modal rumor detection on social media, Neurocomputing 569 (2024) 127112.

[51] Q. Luo, D. Yu, A.M.V.V. Sai, Z. Cai, X. Cheng, A survey of structural representation learning for social networks, Neurocomputing 496 (2022) 56–71.

[52] Y. Zeng, Z. Li, Z. Tang, Z. Chen, H. Ma, Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis, Expert Syst. Appl. 213 (2023) 119240.

[53] Y. Zeng, Z. Li, Z. Chen, H. Ma, Aspect-level sentiment analysis based on semantic heterogeneous graph convolutional network, Front. Comput. Sci. 17 (6) (2023) 176340.

[54] E.M. Mercha, H. Benbrahim, Machine learning and deep learning for sentiment analysis across languages: A survey, Neurocomputing 531 (2023) 195–216.

[55] B. Yang, B. Shao, L. Wu, X. Lin, Multimodal sentiment analysis with unidirectional modality translation, Neurocomputing 467 (2022) 130–137.
[56] S. Moon, L. Neves, V. Carvalho, Multimodal named entity recognition for short social media posts, 2018, arXiv preprint arXiv:1802.07862.
[57] G. Zhao, G. Dong, Y. Shi, H. Yan, W. Xu, S. Li, Entity-level interaction via heterogeneous graph for multimodal named entity recognition, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 6345–6350.
[58] J. Yu, J. Jiang, L. Yang, R. Xia, Improving multimodal named entity recognition via entity span detection with unified multimodal transformer, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3342–3352.
[59] L. Sun, J. Wang, Y. Su, F. Weng, Y. Sun, Z. Zheng, Y. Chen, RIVA: a pre-trained tweet multimodal model based on text-image relation for multimodal NER, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1852–1862.
[60] L. Sun, J. Wang, K. Zhang, Y. Su, F. Weng, RpBERT: a text-image relation propagation-based BERT model for multimodal NER, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 13860–13868.
[61] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, F. Huang, K. Tu, ITA: Image-text alignments for multi-modal named entity recognition, 2021, arXiv preprint arXiv:2112.06482.
[62] M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, J. Li, Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 8032–8040.
[63] X. Wang, J. Cai, Y. Jiang, P. Xie, K. Tu, W. Lu, Named entity and relation extraction with multi-modal retrieval, 2022, arXiv preprint arXiv:2212.01612.
[64] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, G. Zhou, Multi-modal graph fusion for named entity recognition with targeted visual guidance, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 14347–14355.
[65] S. Cui, J. Cao, X. Cong, J. Sheng, Q. Li, T. Liu, J. Shi, Enhancing multimodal entity and relation extraction with variational information bottleneck, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024) 1274–1285.
[66] X. Wang, J. Tian, M. Gui, Z. Li, J. Ye, M. Yan, Y. Xiao, Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition, in: Proceedings of the International Conference on Database Systems for Advanced Applications, Springer, 2022, pp. 297–305.
[67] X. Hu, J. Chen, A. Liu, S. Meng, L. Wen, P.S. Yu, Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5185–5194.
[68] X. Hu, Z. Guo, Z. Teng, I. King, P.S. Yu, Multimodal relation extraction with cross-modal retrieval and synthesis, 2023, arXiv preprint arXiv:2305.16166.
[69] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 1–35.
[70] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, 2021, arXiv preprint arXiv:2104.08691.
[71] Y. Gu, X. Han, Z. Liu, M. Huang, Ppt: Pre-trained prompt tuning for few-shot learning, 2021, arXiv preprint arXiv:2109.04332.
[72] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, 2021, arXiv preprint arXiv:2101.00190.
[73] X. Liu, K. Ji, Y. Fu, W.L. Tam, Z. Du, Z. Yang, J. Tang, P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks, 2021, arXiv preprint arXiv:2110.07602.
[74] M.U. Khattak, H. Rasheed, M. Maaz, S. Khan, F.S. Khan, Maple: Multi-modal prompt learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19113–19122.
[75] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, Y. Zhang, Dual modality prompt tuning for vision-language pre-trained model, IEEE Trans. Multimed. 26 (2023) 2056–2068.
[76] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.
[77] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, J. Luo, A fast and accurate one-stage approach to visual grounding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4683–4693.
[78] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, W. Zhang, Fashion retrieval via graph reasoning networks on a similarity pyramid, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3066–3075.
[79] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016, arXiv preprint arXiv:1603.01354.
[80] D. Zeng, K. Liu, Y. Chen, J. Zhao, Distant supervision for relation extraction via piecewise convolutional neural networks, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1753–1762.
[81] B. Xu, S. Huang, C. Sha, H. Wang, MAF: A general matching and alignment framework for multimodal named entity recognition, in: Proceedings of the 15th ACM International Conference on Web Search and Data Mining, 2022, pp. 1215–1223.
[82] F. Chen, J. Liu, K. Ji, W. Ren, J. Wang, J. Chen, Learning implicit entity-object relations by bidirectional generative alignment for multimodal ner, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4555–4563.
[83] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv preprint arXiv:1908.03557.
[84] F. Chen, Y. Feng, Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction, 2023, arXiv preprint arXiv:2306.14122.
[85] L. Sun, K. Zhang, Q. Li, R. Lou, Umie: Unified multimodal information extraction with instruction tuning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 19062–19070.

Zhenbin Chen is currently studying for a master's degree in the School of Computer Science and Engineering at Guangxi Normal University. His research interests include relation extraction, multi-modal learning and large language model.

Zhixin Li is a professor at School of Computer Science and Engineering, Guangxi Normal University, P. R. China. In 2010, he obtained his Ph.D. degree in computer software and theory from Institute of Computing Technology, Chinese Academy of Sciences. He obtained his B.S. degree and M.S. degree at the Huazhong University of Science and Technology in 1992 and 2004 respectively. His research interests include image understanding, machine learning and cross-media computing. He has won the best doctoral dissertation award of Chinese Association of Artificial Intelligence in 2011.

Mingqi Liu is currently studying for a master's degree in the School of Computer Science and Engineering at Guangxi Normal University. His research interests include sentiment analysis, multimodal learning and large language model.

Canlong Zhang is a professor at School of Computer Science and Engineering, Guangxi Normal University, P. R. China. In 2014, he obtained his Ph.D. degree in control technology and control engineering from Shanghai Jiao Tong University in China. His research interests include target tracking, person re-identification and multi-sensor data fusion.

Huifang Ma is a professor in the College of Computer Science and Engineering at Northwest Normal University, P. R. China. She received the B.E. degree from Northwest Normal University, P. R. China., in 2003, and the M.S. degree from Beijing Normal University, P. R. China., in 2006. She received the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, in 2010. Her research interests include data mining and machine learning.